# New Support Vector Algorithms

Bernhard Schölkopf, GMD [1]  Alex J. Smola, GMD [2]
Robert Williamson, Australian National University [3]
Peter Bartlett, Australian National University [4]

NeuroCOLT2 Technical Report Series

NC2-TR-1998-031

November, 1998[5]

Produced as part of the ESPRIT Working Group
in Neural and Computational Learning II,
NeuroCOLT2 27150

For more information see the NeuroCOLT website
`http://www.neurocolt.com`
or email `neurocolt@neurocolt.com`

[1]bs@first.gmd.de GMD FIRST, Rudower Chaussee 5, 12489 Berlin, Germany

[2]smola@first.gmd.de GMD FIRST, Rudower Chaussee 5, 12489 Berlin, Germany

[3]Robert.Williamson@anu.edu.au, Austrailan National University, Department of Engineering, Canberra 0200, Australia

[4]Peter.Bartlett@anu.edu.au, Austrailan National University, RSISE, Canberra 0200, Australia

[5]Received 26-NOV-98

**Abstract**

We describe a new class of Support Vector algorithms for regression and classification. In these algorithms, a parameter $\nu$ lets one effectively control the number of Support Vectors. While this can be useful in its own right, the parametrization has the additional benefit of enabling us to eliminate one of the other free parameters of the algorithm: the accuracy parameter $\varepsilon$ in the regression case, and the regularization constant $C$ in the classification case.

We describe the algorithms, give some theoretical results concerning the meaning and the choice of $\nu$, and report experimental results.

# 1 Introduction

Support Vector (SV) machines comprise a new class of learning algorithms, motivated by results of statistical learning theory (Vapnik, 1995). Originally developed for pattern recognition (Vapnik and Chervonenkis, 1974; Boser et al., 1992), they represent the decision boundary in terms of a typically small subset (Schölkopf et al., 1995) of all training examples, called the Support Vectors. In order for this property to carry over to the case of SV Regression, Vapnik devised the so-called $\varepsilon$-insensitive loss function

$$|y - f(\mathbf{x})|_\varepsilon = \max\{0, |y - f(\mathbf{x})| - \varepsilon\}, \tag{1}$$

which does not penalize errors below some $\varepsilon > 0$, chosen a priori. His algorithm, which we will henceforth call $\varepsilon$-**SVR**, seeks to estimate functions

$$f(\mathbf{x}) = (\mathbf{w} \cdot \mathbf{x}) + b, \qquad \mathbf{w}, \mathbf{x} \in \mathbf{R}^N, b \in \mathbf{R}, \tag{2}$$

based on independent identically distributed (iid) data

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell) \in \mathcal{X} \times \mathbf{R}. \tag{3}$$

Here, $\mathcal{X}$ is the space in which the input patterns live (e.g., for vectorial data, $\mathcal{X} = \mathbf{R}^N$). The goal of the learning process is to find a function $f$ with a small risk (or test error)

$$R[f] = \int_{\mathcal{X}} l(f, \mathbf{x}, y)\, dP(\mathbf{x}, y), \tag{4}$$

where $P$ is the probability measure which is assumed to be responsible for the generation of the observations (3), and $l$ is a loss function, e.g. $l(f, \mathbf{x}, y) = (f(\mathbf{x}) - y)^2$, or many other choices (Smola and Schölkopf, 1998). The particular loss function for which we would like to minimize (4) depends on the specific regression estimation problem at hand. Note that this does not necessarily have to coincide with the loss function used in our learning algorithm. First, there might be additional constraints that we would like our regression estimation to satisfy, e.g. that it have a sparse representation in terms of the training data — in the SV case, this is achieved through the insensitive zone in (1). Second, we cannot minimize (4) directly in the first place, since we do not know $P$. Instead, we are given the sample (3), and we try to obtain a small risk by minimizing the regularized risk functional

$$\frac{1}{2}\|\mathbf{w}\|^2 + C \cdot R^{\varepsilon}_{emp}[f]. \tag{5}$$

Here, $\|\mathbf{w}\|^2$ is a term which characterizes the model complexity,

$$R^{\varepsilon}_{emp}[f] := \frac{1}{\ell}\sum_{i=1}^{\ell} |y_i - f(\mathbf{x}_i)|_{\varepsilon} \tag{6}$$

measures the $\varepsilon$-insensitive training error, and $C$ is a constant determining the trade-off. In short, minimizing (5) captures the main insight of statistical learning theory, stating that in order to obtain a small risk, one needs to control both training error and model complexity, i.e. explain the data with a simple model.

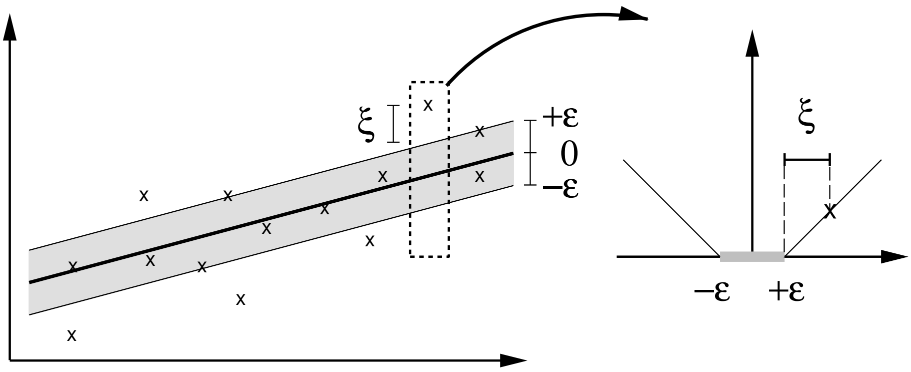


**Figure 1** In SV regression, a desired accuracy $\varepsilon$ is specified a priori. It is then attempted to fit a tube with radius $\varepsilon$ to the data. The trade-off between model complexity and points lying outside of the tube (with positive slack variables $\xi$) is determined by minimizing (5).

The minimization of (5) is equivalent to the following constrained optimization problem (Fig. 1):

$$\text{minimize} \qquad \tau(\mathbf{w}, \boldsymbol{\xi}^{(*)}) = \frac{1}{2}\|\mathbf{w}\|^2 + C \cdot \frac{1}{\ell}\sum_{i=1}^{\ell}(\xi_i + \xi_i^*)) \tag{7}$$

subject to

$$\text{subject to} \qquad ((\mathbf{w}\cdot\mathbf{x}_i) + b) - y_i \;\leq\; \varepsilon + \xi_i \tag{8}$$
$$y_i - ((\mathbf{w}\cdot\mathbf{x}_i) + b) \;\leq\; \varepsilon + \xi_i^* \tag{9}$$
$$\xi_i^{(*)} \;\geq\; 0. \tag{10}$$

Here and below, it is understood that $i = 1, \ldots, \ell$, and that bold face greek letters denote $\ell$-dimensional vectors of the corresponding variables; $^{(*)}$ is a shorthand implying both the variables with and without asterisks.

By using Lagrange multiplier techniques, one can show (Vapnik, 1995) that this leads to the following dual optimization problem. Maximize

$$W(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*) = -\varepsilon \sum_{i=1}^{\ell} (\alpha_i^* + \alpha_i) + \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_i) y_i - \frac{1}{2} \sum_{i,j=1}^{\ell} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(\mathbf{x}_i \cdot \mathbf{x}_j) \tag{11}$$

$$\text{subject to} \qquad \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) = 0 \tag{12}$$

$$\alpha_i^{(*)} \in \left[0, \tfrac{C}{\ell}\right]. \tag{13}$$

The resulting regression estimates are linear; however, the setting can be generalized to a nonlinear one by using the kernel method. As we will use precisely this method in the next section, we shall omit its exposition at this point.

To motivate the new algorithm that we shall propose below, note that the parameter $\varepsilon$ *can* be useful if the desired accuracy of the approximation can be specified beforehand. In some cases, however, we just want the estimate to be as accurate as possible, without having to commit ourselves to a specific level of accuracy a priori.

In the present work, we describe a modification of the $\varepsilon$-SVR algorithm, called $\nu$-SVR, which automatically minimizes $\varepsilon$. Following this, we present two theoretical results on $\nu$-SVR, concerning the connection to robust estimators (section 3) and the asymptotically optimal choice of the parameter $\nu$ (section 4). Next, we extend the algorithm to handle parametric insensitivity models which allow taking into account prior knowledge about heteroscedasticity of the noise. As a bridge connecting this first theoretical part of the paper to the second one, we then present a definition of a *margin* that both SV classification and SV regression algorithms maximize (section 6). In view of this close connection between both algorithms, it is not surprising that it is possible to formulate also a $\nu$-SV classification algorithm. This is done, including some theoretical analysis, in section 7. We conclude with experiments and a discussion.

## 2 $\nu$-SV Regression

To estimate functions (2) from empirical data (3) we proceed as follows (Schölkopf et al., 1998b). At each point $\mathbf{x}_i$, we allow an error of $\varepsilon$. Everything above $\varepsilon$ is captured in slack variables $\xi_i^{(*)}$, which are penalized in the objective function via a regularization constant $C$, chosen a priori (Vapnik, 1995). The size of $\varepsilon$ is traded off against model complexity and slack variables via a constant $\nu \geq 0$:

$$\text{minimize} \qquad \tau(\mathbf{w}, \boldsymbol{\xi}^{(*)}, \varepsilon) = \frac{1}{2}\|\mathbf{w}\|^2 + C \cdot \left(\nu\varepsilon + \frac{1}{\ell}\sum_{i=1}^{\ell}(\xi_i + \xi_i^*)\right) \tag{14}$$

$$\text{subject to} \qquad ((\mathbf{w} \cdot \mathbf{x}_i) + b) - y_i \leq \varepsilon + \xi_i \tag{15}$$

$$y_i - ((\mathbf{w} \cdot \mathbf{x}_i) + b) \leq \varepsilon + \xi_i^* \tag{16}$$

$$\xi_i^{(*)} \geq 0, \quad \varepsilon \geq 0. \tag{17}$$

For the constraints, we introduce multipliers $\alpha_i^{(*)}, \eta_i^{(*)}, \beta \geq 0$, and obtain the Lagrangian

$$L(\mathbf{w}, b, \boldsymbol{\alpha}^{(*)}, \beta, \boldsymbol{\xi}^{(*)}, \varepsilon, \boldsymbol{\eta}^{(*)}) = \tag{18}$$

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\nu\varepsilon + \frac{C}{\ell}\sum_{i=1}^{\ell}(\xi_i + \xi_i^*) - \beta\varepsilon - \sum_{i=1}^{\ell}(\eta_i\xi_i + \eta_i^*\xi_i^*)$$

$$-\sum_{i=1}^{\ell}\alpha_i(\xi_i + y_i - (\mathbf{w}\cdot\mathbf{x}_i) - b + \varepsilon) - \sum_{i=1}^{\ell}\alpha_i^*(\xi_i^* + (\mathbf{w}\cdot\mathbf{x}_i) + b - y_i + \varepsilon)$$

To minimize (14), we have to find the saddle point of $L$, i.e. minimize over the primal variables $\mathbf{w}, \varepsilon, b, \xi_i^{(*)}$ and maximize over the dual variables $\alpha_i^{(*)}, \beta, \eta_i^{(*)}$. Setting the derivatives with respect to the primal variables equal to zero yields the four equations

$$\mathbf{w} = \sum_i(\alpha_i^* - \alpha_i)\mathbf{x}_i \tag{19}$$

$$C\cdot\nu - \sum_i(\alpha_i + \alpha_i^*) - \beta = 0 \tag{20}$$

$$\sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*) = 0 \tag{21}$$

$$\tfrac{C}{\ell} - \alpha_i^{(*)} - \eta_i^{(*)} = 0. \tag{22}$$

In the *SV expansion* (19) only those $\alpha_i^{(*)}$ will be nonzero that correspond to a constraint (15)/(16) which is precisely met. This is due to the Karush-Kuhn-Tucker (KKT) conditions which apply to convex constrained optimization problems (Bertsekas, 1995, e.g.): if we write the constraints as $g(\mathbf{x}_i, y_i) \geq 0$, with corresponding Lagrange multipliers $\alpha_i$, then the point of the solution satisfies $\alpha_i \cdot g(\mathbf{x}_i, y_i) = 0$ for all $i$.

Substituting the above four conditions into $L$ leads to another optimization problem, called the Wolfe dual. Before stating it explicitly, we carry out one further modification. As in Boser et al. (1992), we substitute a kernel $k$ for the dot product, corresponding to a dot product in some feature space related to input space via a nonlinear map $\Phi$,

$$k(\mathbf{x}, \mathbf{y}) = (\Phi(\mathbf{x}) \cdot \Phi(\mathbf{y})). \tag{23}$$

By using $k$, we implicitly carry out all computations in the *feature space* that $\Phi$ maps into — which can have a very high dimensionality. The feature space has the structure of a reproducing kernel Hilbert space (Wahba, 1999; Girosi, 1998; Schölkopf, 1997, e.g.) and hence minimization of $\|\mathbf{w}\|^2$ can be understood in the context of regularization operators (Smola et al., 1998b).

The method is applicable whenever an algorithm can be cast in terms of dot products (Aizerman et al., 1964; Boser et al., 1992; Schölkopf et al., 1998c). The choice of $k$ is a research topic in its own right that we shall not touch in the current essay (Williamson et al., 1998, e.g.); typical choices include Gaussian kernels, $k(\mathbf{x},\mathbf{y}) = \exp(-\|\mathbf{x}-\mathbf{y}\|^2/(2\sigma^2))$, and polynomial kernels, $k(\mathbf{x},\mathbf{y}) = (\mathbf{x}\cdot\mathbf{y})^d$ $(\sigma > 0, d \in \mathbb{N})$.

Rewriting the constraints, noting that $\beta, \eta_i^{(*)} \geq 0$ do not appear in the dual, we arrive at the $\nu$-**SVR Optimization Problem:** for $\nu \geq 0, C > 0$,

$$\text{maximize}\ \ W(\boldsymbol{\alpha}^{(*)}) = \sum_{i=1}^{\ell}(\alpha_i^* - \alpha_i)y_i - \frac{1}{2}\sum_{i,j=1}^{\ell}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)k(\mathbf{x}_i,\mathbf{x}_j) \quad (24)$$

$$\text{subject to} \qquad \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*) = 0 \qquad (25)$$

$$\alpha_i^{(*)} \in \left[0, \tfrac{C}{\ell}\right] \qquad (26)$$

$$\sum_{i=1}^{\ell}(\alpha_i + \alpha_i^*) \leq C \cdot \nu. \qquad (27)$$

The regression estimate then takes the form (cf. (2), (19), (23))

$$f(\mathbf{x}) = \sum_{i=1}^{\ell}(\alpha_i^* - \alpha_i)k(\mathbf{x}_i,\mathbf{x}) + b, \qquad (28)$$

where $b$ (and $\varepsilon$) can be computed by taking into account that (15) and (16) (substitution of $\sum_j(\alpha_j^* - \alpha_j)k(\mathbf{x}_j,\mathbf{x})$ for $(\mathbf{w}\cdot\mathbf{x})$ is understood, cf. (19), (23)) become equalities with $\xi_i^{(*)} = 0$ for points with $0 < \alpha_i^{(*)} < C/\ell$, respectively, due to the KKT conditions.

Before we give theoretical results explaining the significance of the parameter $\nu$, the following observation concerning $\varepsilon$ is helpful. If $\nu > 1$, then $\varepsilon = 0$, since it does not pay to increase $\varepsilon$ (cf. (14)). If $\nu \leq 1$, it can still happen that $\varepsilon = 0$, e.g. if the data are noise-free and can be perfectly interpolated with a low capacity model. The case $\varepsilon = 0$, however, is not what we are interested in; it corresponds to plain $L_1$ loss regression.

Below, we will use the term **errors** to refer to training points lying outside of the tube,[1] and the term **fraction** of errors/SVs to denote the relative numbers of errors/SVs, i.e. divided by $\ell$.

**Proposition 1** *Suppose $\nu$-SVR is applied to some data set and the resulting $\varepsilon > 0$. The following statements hold:*

*(i) $\nu$ is an upper bound on the fraction of errors.*

*(ii) $\nu$ is a lower bound on the fraction of SVs.*

*(iii) Suppose the data (3) were generated iid from a distribution $P(\mathbf{x},y) = P(\mathbf{x})P(y|\mathbf{x})$ with $P(y|\mathbf{x})$ continuous. With probability 1, asymptotically, $\nu$ equals both the fraction of SVs and the fraction of errors.*

[1] For $N > 1$, the "tube" should actually be called a slab, i.e. the region between two parallel hyperplanes.

**Proof**

**Ad (i):** The constraints (26) and (27) imply that at most a fraction $\nu$ of all examples can have $\alpha_i^{(*)} = C/\ell$. All examples with $\xi_i^{(*)} > 0$, i.e. those outside the tube, do certainly satisfy $\alpha_i^{(*)} = C/\ell$ (if not, $\alpha_i^{(*)}$ could grow further to reduce $\xi_i^{(*)}$).

**Ad (ii):** By the KKT conditions, $\varepsilon > 0$ implies $\beta = 0$. Hence, (27) becomes an equality (cf. (20)).[2] Since SVs are those examples for which $0 < \alpha_i^{(*)} \leq C/\ell$ there must be at least $M$ SVs where $MC/\ell = c\nu$ and so $M = \nu\ell$. The result follows (using $\alpha_i \cdot \alpha_i^* = 0$ for all $i$ (Vapnik, 1995)).

**Ad (iii):** The strategy of proof is to show that asymptotically, the probability of a point lying *on* the margin vanishes. Continuity of the conditional distribution $P(y|\mathbf{x})$ implies that for all $f$, and all $t \in \mathbf{R}$, $\lim_{\gamma\to 0} P(|f(\mathbf{x}) - y + t| < \gamma) = 0$. Since the class of SV regression estimates $f$ has well-behaved covering numbers (e.g. Vapnik, 1995; Williamson et al., 1998), we get uniform convergence, so for all $\gamma > 0$, $\sup_f |P(|f(\mathbf{x}) - y + t| < \gamma) - \hat{P}_\ell(|f(\mathbf{x}) - y + t| < \gamma)|$ converges to zero in probability, where $\hat{P}_\ell$ is the sample-based estimate of $P$ (that is, the proportion of points that satisfy $|f(\mathbf{x}) - y + t| < \gamma$). But then for all $\alpha > 0$, $\lim_{\gamma\to 0} \lim_{\ell\to\infty} P(\sup_f \hat{P}_\ell(|f(\mathbf{x}) - y + t| < \gamma) > \alpha) = 0$. Hence, $\sup_f \hat{P}_\ell(|f(\mathbf{x}) - y + t| = 0)$ converges to zero in probability. Thus, the fraction of points with a fixed distance $t$ (in particular, $t = \pm\varepsilon$) to $f$ almost surely converges to 0. Hence the fraction of SVs equals that of errors. Combining (i) and (ii) then shows that both fractions converge almost surely to $\nu$.

∎

Hence, $0 \leq \nu \leq 1$ can be used to control the number of errors (note that for $\nu \geq 1$, (26) implies (27), since $\alpha_i \cdot \alpha_i^* = 0$ for all $i$ (Vapnik, 1995)). Moreover, since the constraint (25) implies that (27) is equivalent to $\sum_i \alpha_i^{(*)} \leq C\nu/2$, we conclude that Proposition 1 actually holds for the upper and the lower edge of the tube separately, with $\nu/2$ each. As an aside, note that by the same argument, the number of SVs at the two edges of the standard $\varepsilon$-SVR tube asymptotically agree.

A more intuitive explanation can be given in terms of the primal objective function (14). At the point of the solution, note that if $\varepsilon > 0$, we must have $(\partial/\partial\varepsilon)\tau(\mathbf{w}, \varepsilon) = 0$, i.e. $\nu + (\partial/\partial\varepsilon)R_{emp}^{\varepsilon} = 0$, hence $\nu = -(\partial/\partial\varepsilon)R_{emp}^{\varepsilon}$. This is greater or equal to the fraction of errors, since the points outside the tube certainly do contribute to a change in $R_{emp}$ when $\varepsilon$ is changed. Possibly, also points at the *edge* of the tube do contribute — this is where the inequality comes from.

Note that this does not contradict our freedom to choose $\nu > 1$. In that case, $\varepsilon = 0$, since it does not pay to increase $\varepsilon$ (cf. (14)).

[2] In practice, one can alternatively work with (27) as an equality constraint, provided that $\nu$ is chosen small enough ($\nu < 1$) to ensure that it does not pay to make $\varepsilon$ negative.

Let us briefly discuss how $\nu$-SVR relates to $\varepsilon$-SVR (section 1). Comparing (11) (substitution of a kernel for the dot product is understood) and (24), we note that $\varepsilon$-SVR requires an additional term $-\varepsilon \sum_{i=1}^{\ell}(\alpha_i^* + \alpha_i)$, which, for fixed $\varepsilon > 0$, encourages that some of the $\alpha_i^{(*)}$ will turn out to be 0. Accordingly, the constraint (27), which appears in $\nu$-SVR, is not needed. The primal problems (7) and (14) differ in the term $\nu\varepsilon$. If $\nu = 0$, then the optimization can grow $\varepsilon$ arbitrarily large, hence zero empirical risk can be otained even when all $\alpha$s are zero.

In the following sense, $\nu$-SVR includes $\varepsilon$-SVR. Note that in the general case, using kernels, $\bar{\mathbf{w}}$ is a vector in feature space.

**Proposition 2** *If $\nu$-SVR leads to the solution $\bar{\varepsilon}, \bar{\mathbf{w}}, \bar{b}$, then $\varepsilon$-SVR with $\varepsilon$ set a priori to $\bar{\varepsilon}$, and the same value of $C$, has the solution $\bar{\mathbf{w}}, \bar{b}$.*

**Proof** If we minimize (14), then fix $\varepsilon$ and minimize only over the remaining variables, the solution does not change. ∎

# 3 The Connection to Robust Estimators

Using the $\varepsilon$-insensitive loss function, only the patterns outside of the $\varepsilon$-tube enter the empirical risk term, whereas the patterns closest to the actual regression have zero loss. This, however, does not mean that it is only the 'outliers' that determine the regression. In fact, the contrary is the case.

**Proposition 3** *Using Support Vector Regression with the $\varepsilon$-insensitive loss function (1), local movements of target values of points outside the tube do not influence the regression.*

**Proof** Shifting $y_i$ locally does not change the status of $(\mathbf{x}_i, y_i)$ as being a point outside the tube. Then the dual solution $\boldsymbol{\alpha}^{(*)}$ is still feasible, i.e. satisfies the constraints (the point still has $\alpha_i^{(*)} = C/\ell$). Moreover, the primal solution, with $\xi_i$ transformed according to the movement of $\mathbf{x}_i$, is also feasible. Finally, the KKT conditions are still satisfied, as still $\alpha_i^{(*)} = C/\ell$. Thus (Bertsekas, 1995, e.g.), $\boldsymbol{\alpha}^{(*)}$ is still the optimal solution. ∎

The proof relies on the fact that everywhere outside the tube, the upper bound on the $\alpha_i^{(*)}$ is the same. This, in turn, is precisely the case if the loss function increases linearly ouside the $\varepsilon$-tube (cf. e.g. Huber (1981) for requirements on robust cost functions). Inside, we could use various functions, with a derivative smaller than the one of the linear part.

For the case of the $\varepsilon$-insensitive loss, the above proposition implies that essentially, the regression is a generalization of an estimator for the mean of a random variable.

(a) It throws away the largest and smallest examples (a fraction $\nu/2$ of either category — in section 2, it is shown that the sum constraint (25) implies that Proposition 1 can be applied separately for the two sides, using $\nu/2$).

(b) It estimates the mean by taking the average of the two extremal ones of the remaining examples.

This is close in spirit to robust estimators like the *trimmed mean.* In fact, we can get closer to the idea of the trimmed mean, which first throws away the largest and smallest points and then computes the mean of the remaining points, by using a quadratic loss inside the $\varepsilon$-tube. This will leave us with Huber's robust loss function.

Finally, we note that the parameter $\nu$ is related to the breakdown point of the corresponding robust estimator (Huber, 1981). As it specifies the fraction of points which may be arbitrarily bad outliers, $\nu$ is related to the fraction of some arbitrary distribution that may be added to a known noise model without the estimator failing.

# 4 Asymptotically Optimal Choice of $\nu$

Using the analysis of Smola et al. (1998a), we can give the optimal $\nu$ for a given class of noise models in the sense of maximizing the statistical efficiency.[3]

**Proposition 4** *Denote $\mathfrak{p}$ a density with unit variance,[4] and $\mathfrak{P}$ a family of noise models generated from $\mathfrak{p}$ by $\mathfrak{P} := \left\{p \,\middle|\, p = \frac{1}{\sigma}\mathfrak{p}\left(\frac{y}{\sigma}\right), \sigma > 0\right\}$. Moreover assume that the data were generated iid from a distribution $p(x,y) = p(x)p(y - f(x))$ with $p(y - f(x))$ continuous. Under the assumption that SV regression produces an estimate $\hat{f}$ converging to the underlying functional dependency $f$, the asymptotically optimal $\nu$, for the estimation-of-location-parameter model of SV regression described in (Smola et al., 1998a), is*

$$\nu = 1 - \int_{-\varepsilon}^{\varepsilon} \mathfrak{p}(t)dt \text{ where } \varepsilon := \operatorname*{argmin}_{\tau} \frac{1}{(\mathfrak{p}(-\tau) + \mathfrak{p}(\tau))^2}\left(1 - \int_{-\tau}^{\tau} \mathfrak{p}(t)dt\right) \quad (29)$$

**Proof** Under the assumptions stated above, the probability of a deviation larger than $\varepsilon$, $Pr\left\{|y - \hat{f}(x)| > \varepsilon\right\}$, will converge to

$$Pr\left\{|y - f(x)| > \varepsilon\right\} = \int_{\mathfrak{X}\times\{\mathbb{R}\setminus[-\varepsilon,\varepsilon]\}} p(x)p(\xi)dxd\xi = 1 - \int_{-\varepsilon}^{\varepsilon} p(\xi)d\xi. \quad (30)$$

This is also the fraction of samples that will (asymptotically) become SVs (Proposition 1, (iii)). Therefore an algorithm generating a fraction $\nu = 1 - \int_{-\varepsilon}^{\varepsilon} p(\xi)d\xi$ SVs will correspond to an algorithm with a tube of size $\varepsilon$. The consequence is that, given a noise model $p(\xi)$, one can compute the optimal $\varepsilon$ for it, and then, by using (30), compute the corresponding optimal value $\nu$. Next one exploits the linear scaling behaviour between the standard deviation $\sigma$ of a distribution $p$ and the optimal $\varepsilon$, established in (Smola et al., 1998a). This

[3] This section assumes familiarity with some concepts of information geometry. A more complete explanation of the model underlying the argument is given in Smola et al. (1998a) and can be downloaded from http://svm.first.gmd.de.

[4] $\mathfrak{p}$ is just a prototype generating the class of densities $\mathfrak{P}$. Normalization assumptions are made merely for the ease of notation.

means that one only has to consider distributions of unit variance, say, $\mathfrak{p}$, to compute an optimal value of $\nu$ that holds for the whole class of distributions $\mathfrak{P}$. The last step is to compute the estimator efficiency as $e\left(\frac{\varepsilon}{\sigma}\right) = \frac{Q^2}{GI}$ (written in information geometry notation; cf. Smola et al. (1998a)), where the Fisher information $I$ is independent of $\varepsilon$, and thus by (Smola et al., 1998a)

$$\frac{1}{e(\varepsilon)} \propto \frac{Q^2}{G} = \frac{1}{(\mathfrak{p}(-\varepsilon) + \mathfrak{p}(\varepsilon))^2}\left(1 - \int_{-\varepsilon}^{\varepsilon} \mathfrak{p}(t)dt\right) \tag{31}$$

The minimum of (31) yields the optimal choice of $\varepsilon$, which allows computation of the corresponding $\nu$ and thus proves the proposition. ∎

Consider now an example: arbitrary polynomial noise models ($\propto e^{-|\xi|^p}$) with unit variance can be written as

$$\mathfrak{p}(\xi) = \frac{1}{2}\sqrt{\frac{\Gamma(3/p)}{\Gamma(1/p)}}\frac{p}{\Gamma(1/p)}\exp\left(-\left(\sqrt{\frac{\Gamma(3/p)}{\Gamma(1/p)}}|\xi|\right)^p\right) \tag{32}$$

where $\Gamma$ denotes the gamma function. Table 1 shows the optimal value of $\nu$ for different polynomial degrees. Observe that the more "lighter-tailed" the

| polynomial degree $p$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| optimal $\nu$ | 1.00 | 0.54 | 0.29 | 0.19 | 0.14 | 0.11 | 0.09 | 0.07 |

**Table 1** Optimal $\nu$ for various degrees of polynomial additive noise.

distribution becomes, the smaller $\nu$ are optimal, i.e. the tube width increases. This is reasonable as only for very long tails of the distribution (data with many outliers) it appears reasonable to use an early cutoff on the influence of the data (by basically giving all data equal influence via $\alpha_i = C/\ell$). The extreme case of Laplacian noise ($\nu = 1$) leads to a tube width of 0, i.e. to $L_1$ regression.

We conclude this section with three caveats: first, the proposition only makes an asymptotic statement; second, for nonzero $\varepsilon$, the SV regression need not necessarily converge to the target $f$: measured using $|.|_\varepsilon$, many other functions are just as good as $f$ itself; third, the proportionality between $\varepsilon$ and $\sigma$ has only been established in the estimation-of-location-parameter context, which is not quite SV regression.

## 5 Parametric Insensitivity Models

We now return to the algorithm described in section 2. We generalized $\varepsilon$-SVR by considering the tube as not given but instead estimated it as a model parameter. What we have so far retained is the assumption that the $\varepsilon$-insensitive zone has a tube (or slab) shape. We now go one step further and use parametric models of arbitrary shape (Schölkopf et al., 1998a). This can be useful in situations where the noise depends on $\mathbf{x}$, i.e. where it is heteroscedastic.

Let $\{\zeta_q^{(*)}\}$ (here and below, $q = 1, \ldots, p$ is understood) be a set of $2p$ positive functions on the input space $\mathfrak{X}$. Consider the following quadratic program: for given $\nu_1^{(*)}, \ldots, \nu_p^{(*)} \geq 0$, minimize

$$\tau(\mathbf{w}, \boldsymbol{\xi}^{(*)}, \boldsymbol{\varepsilon}^{(*)}) = \|\mathbf{w}\|^2/2 + C \cdot \left( \sum_{q=1}^{p} (\nu_q \varepsilon_q + \nu_q^* \varepsilon_q^*) + \frac{1}{\ell} \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) \right) \tag{33}$$

$$\begin{aligned} \text{subject to} \quad ((\mathbf{w} \cdot \mathbf{x}_i) + b) - y_i &\leq \textstyle\sum_q \varepsilon_q \zeta_q(\mathbf{x}_i) + \xi_i & (34) \\ y_i - ((\mathbf{w} \cdot \mathbf{x}_i) + b) &\leq \textstyle\sum_q \varepsilon_q^* \zeta_q^*(\mathbf{x}_i) + \xi_i^* & (35) \\ \xi_i^{(*)} \geq 0, \quad \varepsilon_q^{(*)} &\geq 0. & (36) \end{aligned}$$

A calculation analogous to that in section 2 shows that the Wolfe dual consists of maximizing (24) subject to (25), (26), and, instead of (27), the modified constraints, still linear in $\boldsymbol{\alpha}^{(*)}$,

$$\sum_{i=1}^{\ell} \alpha_i^{(*)} \zeta_q^{(*)}(\mathbf{x}_i) \leq C \cdot \nu_q^{(*)}. \tag{37}$$

In the experiments in section 8, we use a simplified version of this optimization problem, where we drop the term $\nu_q^* \varepsilon_q^*$ from the objective function (33), and use $\varepsilon_q$ and $\zeta_q$ in (35). By this, we render the problem symmetric with respect to the two edges of the tube. In addition, we use $p = 1$. This leads to the same Wolfe dual, except for the last constraint, which becomes (cf. (27))

$$\sum_{i=1}^{\ell} (\alpha_i + \alpha_i^*) \zeta(\mathbf{x}_i) \leq C \cdot \nu. \tag{38}$$

Note that the optimization problem of section 2 can be recovered by using the constant function $\zeta \equiv 1$.[5]

The advantage of this setting is that since the same $\nu$ is used for both sides of the tube, the computation of $\varepsilon, b$ is straightforward: for instance, by solving a linear system, using two conditions as those described following (28). Otherwise, general statements are harder to make: the linear system can have a zero determinant, depending on whether the functions $\zeta_p^{(*)}$, evaluated on the $\mathbf{x}_i$ with $0 < \alpha_i^{(*)} < C/\ell$, are linearly dependent. The latter occurs, for instance, if we use constant functions $\zeta^{(*)} \equiv 1$. In this case, it is pointless to use two different values $\nu, \nu^*$; for, the constraint (25) then implies that *both* sums $\sum_{i=1}^{\ell} \alpha_i^{(*)}$ will be bounded by $C \cdot \min\{\nu, \nu^*\}$. We conclude this section by giving, without proof, a generalization of Proposition 1 to the optimization problem with constraint (38):

**Proposition 5** *Suppose we run the above algorithm on a data set with the result that $\varepsilon > 0$. Then*

[5]Observe the similarity to semiparametric SV models (Smola et al., 1998d) where a modification of the expansion of $f$ led to similar additional constraints. The important difference in the present setting is that the Lagrange multipliers $\alpha_i$ and $\alpha_i^*$ are treated equally and not with different signs as in semiparametric modelling.

*(i)* $\frac{\nu\ell}{\sum_i \zeta(\mathbf{x}_i)}$ *is an upper bound on the fraction of errors.*

*(ii)* $\frac{\nu\ell}{\sum_i \zeta(\mathbf{x}_i)}$ *is an upper bound on the fraction of SVs.*

*(iii) Suppose the data (3) were generated iid from a distribution* $P(\mathbf{x}, y) = P(\mathbf{x})P(y|\mathbf{x})$ *with* $P(y|\mathbf{x})$ *continuous. With probability 1, asymptotically, the fractions of SVs and errors equal* $\nu \cdot (\int \zeta(\mathbf{x})\, d\tilde{P}(\mathbf{x}))^{-1}$, *where* $\tilde{P}$ *is the asymptotic distribution of SVs over* $\mathbf{x}$.

## 6 Margins in Regression and Classification

Historically, the SV algorithm was first proposed for the case of pattern recognition (Vapnik and Chervonenkis, 1974; Boser et al., 1992), and then generalized to regression (Vapnik, 1995). Conceptually, however, one can take the view that the latter case is actually the simpler one — providing a posterior justification why we started this paper with the regression case. To explain this, we will introduce a suitable definition of a *margin* that is maximized in both cases.

At first glance, the two variants of the algorithm seem conceptually different. In the case of pattern recognition, a margin of separation between two pattern classes is maximized, and the Support Vectors are those examples which lie closest to this margin. In the simplest case, where the training error is fixed to 0, this is done by minimizing $\|\mathbf{w}\|^2$ subject to $y_i \cdot ((\mathbf{w} \cdot \mathbf{x}_i) + b) \geq 1$ (note that in pattern recognition, the targets $y_i$ are in $\{\pm 1\}$).

In regression estimation, on the other hand, a tube of radius $\varepsilon$ is fitted to the data, in the space of the target values, with the property that it corresponds to the flattest function in feature space. Here, the Support Vectors lie at the edge of the tube. The parameter $\varepsilon$ does not occur in the pattern recognition case.

We will presently show how these seemingly different problems are identical, how this naturally leads to the concept of *canonical* hyperplanes (Vapnik, 1995), and how it suggests different generalizations to the estimation of vector-valued functions.

**Definition 6 ($\varepsilon$-margin)** *Let* $(E, \|.\|_E)$, $(F, \|.\|_F)$ *be normed spaces, and* $\mathcal{X} \subset E$. *We define the* $\varepsilon$*-margin of a function* $f : \mathcal{X} \to F$ *as*

$$m_\varepsilon(f) := \inf\{\|\mathbf{x} - \mathbf{y}\|_E : \mathbf{x}, \mathbf{y} \in \mathcal{X}, \|f(\mathbf{x}) - f(\mathbf{y})\|_F \geq 2\varepsilon\}. \tag{39}$$

$m_\varepsilon(f)$ can be zero, even for continuous functions, an example being $f(x) = 1/x$ on $\mathcal{X} = \mathbf{R}^+$. There, $m_\varepsilon(f) = 0$ for all $\varepsilon > 0$.

Note that the $\varepsilon$-margin is related (albeit not identical) to the modulus of continuity: given $\delta > 0$, the latter measures the largest difference in function values which can be obtained using points within a distance $\delta$ in $E$.

The following observations characterize the functions for which the margin is strictly positive.

**Lemma 7 (Uniformly continuous functions)** *With the above notations,* $m_\varepsilon(F)$ *is positive for all* $\varepsilon > 0$ *if and only if* $f$ *is uniformly continuous.*

**Proof** By definition of $m_\varepsilon$, we have

$$\Big(\|f(\mathbf{x})-f(\mathbf{y})\|_F \geq 2\varepsilon \implies \|\mathbf{x}-\mathbf{y}\|_E \geq m_\varepsilon(f)\Big) \tag{40}$$

$$\iff \Big(\|\mathbf{x}-\mathbf{y}\|_E < m_\varepsilon(f) \implies \|f(\mathbf{x})-f(\mathbf{y})\|_F < 2\varepsilon\Big), \tag{41}$$

i.e., if $m_\varepsilon(f) > 0$, then $f$ is uniformly continuous. Similarly, if $f$ is uniformly continuous, then for each $\varepsilon > 0$, we can find a $\delta > 0$ such that $\|f(\mathbf{x})-f(\mathbf{y})\|_F \geq 2\varepsilon$ implies $\|\mathbf{x}-\mathbf{y}\|_E \geq \delta$. Since the latter holds uniformly, we can take the infimum to get $m_\varepsilon(f) \geq \delta > 0$. ∎

We next specialize to a particular set of uniformly continuous functions.

**Lemma 8 (Lipschitz-continuous functions)** *If there exists some $L > 0$ such that for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, $\|f(\mathbf{x})-f(\mathbf{y})\|_F \leq L \cdot \|\mathbf{x}-\mathbf{y}\|_E$, then $m_\varepsilon \geq \frac{2\varepsilon}{L}$.*

**Proof** Take the infimum over $\|\mathbf{x}-\mathbf{y}\|_E \geq \frac{\|f(\mathbf{x})-f(\mathbf{y})\|_F}{L} \geq \frac{2\varepsilon}{L}$. ∎

**Example 9 (SV regression estimation)** *Suppose that $E$ is endowed with a dot product $(\cdot\,\cdot\,\cdot)$ (generating the norm $\|.\|_E$). For linear functions (2), the margin takes the form $m_\varepsilon(f) = \frac{2\varepsilon}{\|\mathbf{w}\|}$. To see this, note that due to $|f(\mathbf{x})-f(\mathbf{y})| = |(\mathbf{w}\cdot(\mathbf{x}-\mathbf{y}))|$, the distance $\|\mathbf{x}-\mathbf{y}\|$ will get smallest if $|(\mathbf{w}\cdot(\mathbf{x}-\mathbf{y}))| = 2\varepsilon$, with $\mathbf{x}-\mathbf{y}$ parallel to $\mathbf{w}$ (due to Cauchy-Schwartz), i.e. if $\mathbf{x}-\mathbf{y} = \pm 2\varepsilon\mathbf{w}/\|\mathbf{w}\|^2$. In that case, $\|\mathbf{x}-\mathbf{y}\| = 2\varepsilon/\|\mathbf{w}\|$. For fixed $\varepsilon > 0$, maximizing the margin hence amounts to minimizing $\|\mathbf{w}\|$, as done in SV regression: in the simplest form, cf. (7) without slack variables $\xi_i$, the training on data (3) consists of minimizing $\|\mathbf{w}\|^2$ subject to*

$$|f(\mathbf{x}_i) - y_i| \leq \varepsilon. \tag{42}$$

**Example 10 (SV pattern recognition, Fig. 2)** *We specialize the setting of Example 9 to the case where $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_\ell\}$. Then $m_1(f) = \frac{2}{\|\mathbf{w}\|}$ is equal to the margin defined for Vapnik's* canonical hyperplane *(Vapnik, 1995). The latter is a way in which, given the data set $\mathcal{X}$, an oriented hyperplane in $E$ can be uniquely expressed by a linear function (2), requiring that*

$$\min\{|f(\mathbf{x})| : \mathbf{x} \in \mathcal{X}\} = 1. \tag{43}$$

*Vapnik gives a bound on the VC-dimension of canonical hyperplanes in terms of $\|\mathbf{w}\|$. An optimal margin SV machine for pattern recognition can be constructed from data*

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell) \in \mathcal{X} \times \{\pm 1\} \tag{44}$$

*as follows (Boser et al., 1992):*

$$\textit{minimize} \quad \|\mathbf{w}\|^2 \quad \textit{subject to} \quad y_i \cdot f(\mathbf{x}_i) \geq 1. \tag{45}$$

*The decision function which is used for classification takes the form*

$$f^*(\mathbf{x}) = sgn((\mathbf{w}\cdot\mathbf{x}) + b). \tag{46}$$

*The parameter $\varepsilon$ is superfluous in pattern recognition, as the resulting decision function*

$$f^*(\mathbf{x}) = sgn((\mathbf{w} \cdot \mathbf{x}) + b), \tag{47}$$

*will not change if we minimize $\|\mathbf{w}\|^2$ subject to*

$$y_i \cdot f(\mathbf{x}_i) \geq \varepsilon. \tag{48}$$

*Finally, to understand why the constraint (45) looks different from (42) (e.g., one is multiplicative, the other one additive), note that in regression, the points $(\mathbf{x}_i, y_i)$ are required to lie within a tube of radius $\varepsilon$, whereas in pattern recognition, they are required to lie outside of the tube (Fig. 2), and on the correct side. For the points* on *the tube, we have $1 = y_i \cdot f(\mathbf{x}_i) = 1 - |f(\mathbf{x}_i) - y_i|$.*

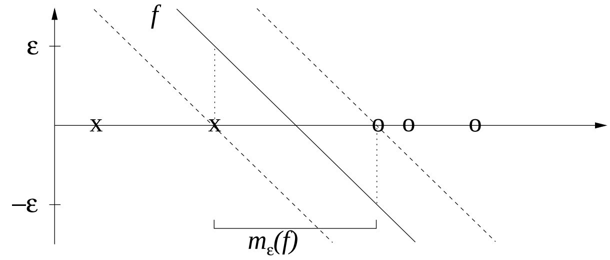


**Figure 2** 1-D toy problem: separate 'x' from 'o'. The SV classification algorithm constructs a linear function $f(x) = w \cdot x + b$ satisfying (43) ($\varepsilon = 1$). To maximize the margin $m_\varepsilon(f)$, one has to minimize $|w|$.

So far, we have only interpreted known algorithms in terms of maximizing $m_\varepsilon$. Next, we consider the question whether we can use the latter as a guide for constructing more general algorithms.

**Example 11 (SV regression for vector-valued functions)** *Assume $E = \mathbf{R}^N$. For linear functions $f(\mathbf{x}) = W\mathbf{x} + \mathbf{b}$, with $W$ being an $N \times N$ matrix, and $\mathbf{b} \in \mathbf{R}^N$, we have, as a consequence of Lemma 8,*

$$m_\varepsilon(f) \geq \frac{2\varepsilon}{\|W\|}, \tag{49}$$

*where $\|W\|$ is any matrix norm of $W$ which is compatible (Horn and Johnson, 1985) with $\|.\|_E$. If the matrix norm is the one induced by $\|.\|_E$, i.e. there exists a unit vector $\mathbf{z} \in E$ such that $\|W\mathbf{z}\|_E = \|W\|$, then equality holds in (49). To see the latter, we use the same argument as in Example 9, setting $\mathbf{x} - \mathbf{y} = 2\varepsilon\mathbf{z}/\|W\|$.*

*For the Hilbert-Schmidt norm $\|W\|_2 = \sqrt{\sum_{i,j=1}^N W_{ij}^2}$, which is compatible with the vector norm $\|.\|_2$, the problem of minimizing $\|W\|$ subject to separate constraints for each output dimension separates into $N$ regression problems.*

*In (Smola et al., 1998c), it is shown that one can specify invariance requirements which imply that the regularizer act on the output dimensions separately and identically (i.e. in a scalar fashion). In particular it turns out that under the assumption of quadratic homogenity and permutation symmetry, the Hilbert–Schmidt norm is the only admissible one.*

# 7 $\nu$-SV Classification

We saw that $\nu$-SVR differs from $\varepsilon$-SVR in that it uses the parameters $\nu$ and $C$ instead of $\varepsilon$ and $C$. In many cases, this is a useful re-parametrization of the original algorithm, and thus it is worthwhile to ask the question whether a similar change could be incorporated in the original SV classification algorithm (for brevity, let us call it $C$-SVC). There, the primal optimization problem is to minimize (Cortes and Vapnik, 1995)

$$\tau(\mathbf{w}, \boldsymbol{\xi}) = \frac{1}{2}\|\mathbf{w}\|^2 + \tfrac{C}{\ell}\sum_i \xi_i \tag{50}$$

subject to

$$y_i \cdot ((\mathbf{x}_i \cdot \mathbf{w}) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0. \tag{51}$$

The goal of the learning process is to estimate a function $f^*$ (47) such that the probability of misclassification on an independent test set, the risk $R[f^*]$, is small.[6]

Therefore, the only parameter that we can dispose of is the regularization constant $C$. To substitute it by a parameter similar to the $\nu$ used in the regression case, we proceed as follows. As a primal problem for $\nu$-SVC, we consider the minimization of

$$\tau(\mathbf{w}, \boldsymbol{\xi}, \rho) = \frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + \frac{1}{\ell}\sum_i \xi_i \tag{52}$$

subject to (cf. (48))

$$y_i \cdot ((\mathbf{x}_i \cdot \mathbf{w}) + b) \geq \rho - \xi_i, \tag{53}$$

$$\xi_i \geq 0, \quad \rho \geq 0. \tag{54}$$

For reasons we shall expain below, no constant $C$ appears in this formulation. To understand the role of $\rho$, note that for $\boldsymbol{\xi} = 0$, the constraint (53) simply states that the two classes are separated by the *margin* $2\rho/\|\mathbf{w}\|$ (cf. Example 10).

To derive the dual, we consider the Lagrangian

$$\begin{aligned} L(\mathbf{w}, \boldsymbol{\xi}, b, \rho, \boldsymbol{\alpha}, \boldsymbol{\beta}, \delta) \; = \; & \frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + \frac{1}{\ell}\sum_i \xi_i \\ & - \sum_i (\alpha_i(y_i((\mathbf{x}_i \cdot \mathbf{w}) + b) - \rho + \xi_i) - \beta_i \xi_i) - \delta\rho, \end{aligned} \tag{55}$$

using multipliers $\alpha_i, \beta_i, \delta \geq 0$. This function has to be mimimized with respect to the primal variables $\mathbf{w}, \boldsymbol{\xi}, b, \rho$ and maximized with respect to the dual variables $\boldsymbol{\alpha}, \boldsymbol{\beta}, \delta$. To eliminate the former, we compute the corresponding partial derivatives and set them to 0, obtaining the following conditions:

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \tag{56}$$

$$\alpha_i + \beta_i = 1/\ell, \quad 0 = \sum_i \alpha_i y_i, \quad \sum_i \alpha_i - \delta = \nu \tag{57}$$

[6]Implicitly we make use of the $\{0, 1\}$-loss function, hence the risk equals the probability of misclassification.

In the *SV expansion* (56) only those $\alpha_i$ can be nonzero that correspond to a constraint (53) which is precisely met (KKT conditions, cf. Vapnik (1995)).

Substituting (56) and (57) into $L$, using $\alpha_i, \beta_i, \delta \geq 0$, and incorporating kernels for dot products leaves us with the following quadratic optimization problem: maximize

$$W(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{ij} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \tag{58}$$

subject to

$$0 \leq \alpha_i \leq 1/\ell \tag{59}$$

$$0 = \sum_i \alpha_i y_i \tag{60}$$

$$\sum_i \alpha_i \geq \nu \tag{61}$$

The resulting decision function can be shown to take the form

$$f^*(\mathbf{x}) = \operatorname{sgn}\left(\sum_i \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) + b\right). \tag{62}$$

Compared to the original dual (Boser et al., 1992; Vapnik, 1995), there are two differences. First, there is an additional constraint (61), similar to the regression case (27). Second, the linear term $\sum_i \alpha_i$ of Boser et al. (1992) no more appears in the objective function (58). This has an interesting consequence: (58) is now quadratically homogeneous in $\boldsymbol{\alpha}$. It is straightforward to verify that one obtains exactly the same objective function if one starts with the primal function $\tau(\mathbf{w}, \boldsymbol{\xi}, \rho) = \|\mathbf{w}\|^2/2 + C \cdot (-\nu\rho + (1/\ell)\sum_i \xi_i)$ (i.e., if one uses $C$), the only difference being that the constraints (59) and (61) would have an extra factor $C$ on the right hand side. In that case, due to the homogeneity, the solution of the dual would be scaled by $C$; however, it is straightforward to see that the corresponding decision function will not change. Hence we may set $C = 1$.

To compute $b$ and $\rho$, we consider two sets $S_\pm$, of identical size $s > 0$, containing SVs $\mathbf{x}_i$ with $0 < \alpha_i < 1$ and $y_i = \pm 1$, respectively. Then, due to the KKT conditions, (53) becomes an equality with $\xi_i = 0$. Hence, in terms of kernels

$$b = -\frac{1}{2s} \sum_{\mathbf{x} \in S_+ \cup S_-} \sum_j \alpha_j y_j k(\mathbf{x}, \mathbf{x}_j), \tag{63}$$

$$\rho = \frac{1}{2s}\Big( \sum_{\mathbf{x} \in S_+} \sum_j \alpha_j y_j k(\mathbf{x}, \mathbf{x}_j) - \sum_{\mathbf{x} \in S_-} \sum_j \alpha_j y_j k(\mathbf{x}, \mathbf{x}_j)\Big), \tag{64}$$

As in the regression case, the $\nu$ parameter has a more natural interpretation than the one we removed, $C$. To formulate it, let us first define the term **margin error**: by this, we denote points with $\xi_i > 0$, i.e. which are either errors or lie within the margin. Formally, the fraction of margin errors is

$$R^\rho_{emp}[f] := \frac{1}{\ell}|\{i : f(\mathbf{x}_i) < \rho \textit{ or } \operatorname{sgn}(f(\mathbf{x}_i)) \neq y_i\}|. \tag{65}$$

Here, $f$ is used to denote the argument of the sgn in the decision function (62), i.e. $f^* = \operatorname{sgn} \circ f$.

We are now in a position to modify Proposition 1 for the case of pattern recognition:

**Proposition 12** *Suppose $k(\cdot,\cdot)$ is a real analytic kernel function and we run $\nu$-SVC with $k(\cdot,\cdot)$ on some data with the resulting $\rho > 0$. Then*

*(i) $\nu$ is an upper bound on the fraction of margin errors.*

*(ii) $\nu$ is a lower bound on the fraction of SVs.*

*(iii) Suppose the data (44) were generated iid from a distribution $P(\mathbf{x}, y) = P(\mathbf{x})P(y|\mathbf{x})$ such that neither $P(\mathbf{x}, y = 1)$ nor $P(\mathbf{x}, y = -1)$ contains any discrete component. Suppose, moreover, that the kernel is analytic and non-constant. With probability 1, asymptotically, $\nu$ equals both the fraction of SVs and the fraction of errors.*

**Proof**

**Ad (i):** By the KKT conditions, $\rho > 0$ implies $\delta = 0$. Hence, (61) becomes an equality (cf. (57)). Thus, at most a fraction $\nu$ of all examples can have $\alpha_i = 1/\ell$. All examples with $\xi_i > 0$ do satisfy $\alpha_i = 1/\ell$ (if not, $\alpha_i$ could grow further to reduce $\xi_i$).

**Ad (ii):** SVs can contribute at most $1/\ell$ to (61), hence there must be at least $\nu\ell$ of them.

**Ad (iii):** It follows from the condition on $P(\mathbf{x}, y)$ that apart from some set of measure zero (arising from possible singular components), the two class distributions are absolutely continuous and can be written as integrals over distribution functions. As the kernel is analytic and non-constant, it cannot be constant in any open set — otherwise it would be constant everywhere. Therefore, functions $f$ constituting the argument of the sgn in the SV decision function ((62); essentially: the class of SV regression functions) transform the distribution over $\mathbf{x}$ into distributions such that for all $f$, and all $t \in \mathbf{R}$, $\lim_{\gamma\to 0} P(|f(\mathbf{x})+t| < \gamma) = 0$. At the same time, we know that the class of these functions has well-behaved covering numbers, hence we get uniform convergence: for all $\gamma > 0$, $\sup_f |P(|f(\mathbf{x}) + t| < \gamma) - \hat{P}_\ell(|f(\mathbf{x}) + t| < \gamma)|$ converges to zero in probability, where $\hat{P}_\ell$ is the sample-based estimate of $P$ (that is, the proportion of points that satisy $|f(\mathbf{x}) + t| < \gamma$). But then for all $\alpha > 0$, $\lim_{\gamma\to 0} \lim_{\ell\to\infty} P(\sup_f \hat{P}_\ell(|f(\mathbf{x}) + t| < \gamma) > \alpha) = 0$. Hence, $\sup_f \hat{P}_\ell(|f(\mathbf{x}) + t| = 0)$ converges to zero in probability. Using $t = \pm\rho$ thus shows that almost surely the fraction of points exactly on the margin tends to zero, hence the fraction of SVs equals that of margin errors. Combining (i) and (ii) then shows that both fractions converge almost surely to $\nu$. ∎

Moreover, since (60) means that the sums over the coefficients of positive and negative SVs respectively are equal, we conclude that Proposition 12 actually holds for both classes separately, with $\nu/2$. As an aside, note that by the same argument, the number of SVs at the two sides of the margin asymptotically agree.

A connection to standard SV classification, and a somewhat surprising interpretation of the regularization parameter $C$, is described by the following result:

**Proposition 13** *If $\nu$-SV classification leads to $\rho > 0$, then C-SV classification, with C set a priori to $1/\rho$, leads to the same decision function.*

**Proof** If one minimizes (52) and then fixes $\rho$ to minimize only over the remaining variables, nothing will change. Hence the obtained solution $\mathbf{w}_0, b_0, \boldsymbol{\xi}_0$ minimizes (50), for $C = 1$, subject to (53). To recover the constraint (51), we rescale to the set of variables $\mathbf{w}' = \mathbf{w}/\rho, b' = b/\rho, \boldsymbol{\xi}' = \boldsymbol{\xi}/\rho$. This leaves us, up to a constant scaling factor $\rho^2$, with the objective function (50), using $C = 1/\rho$. ∎

For the selection of $C$, several methods have been proposed which could probably be adapted for $\nu$ (Schölkopf, 1997; Shawe-Taylor and Cristianini, 1999, e.g.). In practice, most researchers have so far used cross validation. Clearly, this could be done also for $\nu$-SVC. Nevertheless, we shall presently propose a method which takes into account specific properties of $\nu$-SVC.

The parameter $\nu$ lets us control the number of margin errors, which, in turn, is the crucial quantity in a class of bounds on the the generalization error of classifiers using covering numbers to measure the classifier capacity. We can use this connection to give a generalization error bound for $\nu$-SVC in terms of $\nu$. There are a number of complications in doing this the best possible way, and so here we will simply indicate the simplest result. It is based on the following result:

**Proposition 14 (Bartlett (1998))** *Suppose $\rho > 0, 0 < \delta < \frac{1}{2}$, P is a probability distribution on $\mathcal{X} \times \{-1, 1\}$ from which the training set (44) is drawn. Then with probability at least $1 - \delta$ for every $f$ in some function class $\mathcal{F}$, the probability of error of the classification function $f^* = sgn \circ f$ on an independent test set is bounded according to*

$$R[f^*] \leq R^{\rho}_{emp}[f] + \sqrt{\tfrac{2}{\ell}\left(\ln \mathcal{N}(\mathcal{F}, l_\infty^{2\ell}, \rho/2) + \ln(2/\delta)\right)}, \tag{66}$$

*where $\mathcal{N}(\mathcal{F}, l_\infty^{\ell}, \rho) = \sup_{\mathbf{X}=\mathbf{x}_1,\ldots,\mathbf{x}_\ell} \mathcal{N}(\mathcal{F}|_{\mathbf{X}}, l_\infty, \rho)$, $\mathcal{F}|_{\mathbf{X}} = \{(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_\ell)) \colon f \in \mathcal{F}\}$ and $l_\infty$ is the usual $l_\infty$ metric on a set of vectors.*

To obtain the generalization bound for $\nu$-SVC, we simply substitute the bound $R^{\rho}_{emp}[f] \leq \nu$ (Proposition 12, (i)) and some estimate of the covering numbers in terms of the margin. The best currently available bounds are stated in terms of the functional inverse of $\mathcal{N}$, hence the slightly complicated expressions in the following.

**Proposition 15 (Williamson et al. (1998))** *Denote $B_R$ the ball of radius R around the origin in some Hilbert space F. Then the covering number $\mathcal{N}$ of the class of functions*

$$\mathcal{F} = \{\mathbf{x} \mapsto (\mathbf{w} \cdot \mathbf{x}) : \|\mathbf{w}\| \leq 1, \mathbf{x} \in B_R\} \tag{67}$$

*at scale $\rho$ satisfies*

$$\log_2 \mathcal{N}(\mathcal{F}, l_\infty^{\ell}, \rho) \leq \inf\left\{n \,\middle|\, \tfrac{c^2 B_R^2}{\rho^2} \tfrac{1}{n} \log_2\left(1 + \tfrac{\ell}{n}\right) \geq 1\right\} - 1 \tag{68}$$

*where c is a constant (some numerical estimates are being computed).*

This is a consequence of a fundamental theoren by due to Maurey. For $\ell \geq 2$ one thus obtains

$$\log_2 \mathcal{N}(\mathcal{F}, l_\infty^\ell, \rho) \leq \frac{c^2 B_R^2}{\rho^2} \log_2 \ell - 1 \tag{69}$$

To apply these results to $\nu$-SVC, we rescale $\mathbf{w}$ to length 1, thus obtaining a margin $\rho/\|\mathbf{w}\|$ (cf. (53)). Moreover we have to combine Proposition 14 and Proposition 15. Using $\rho/2$ instead of $\rho$ in the latter, yields the following result.

**Proposition 16** *Suppose $\nu$-SVC is used with a kernel of the form $k(\mathbf{x}, \mathbf{y}) = k(\|\mathbf{x} - \mathbf{y}\|)$ with $k(0) = 1$. Then all the data points $\Phi(\mathbf{x}_i)$ in feature space live in a ball of radius 1 centered at the origin. Consequently with probability at least $1 - \delta$ over the training set (44), the $\nu$-SVC decision function $f^* = sgn \circ f$, with $f(\mathbf{x}) = \sum_i \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i)$ (cf. (62)), has a probability of test error bounded according to*

$$\begin{aligned} R[f^*] &\leq R_{emp}^\rho[f] + \sqrt{\frac{2}{\ell}\left(\frac{4c^2\|\mathbf{w}\|^2}{\rho^2}\log_2(2\ell) - 1 + \ln(2/\delta)\right)} \\ &\leq \nu + \sqrt{\frac{2}{\ell}\left(\frac{4c^2\|\mathbf{w}\|^2}{\rho^2}\log_2(2\ell) - 1 + \ln(2/\delta)\right)} \end{aligned}$$

Notice that in general, $\|\mathbf{w}\|$ is a vector in feature space.

Note that the set of functions in the Proposition differs from Support Vector decision functions (62) in that it comes without the "$+b$" term. This leads to a minor modification, for details see (Williamson et al., 1998).

Better bounds can be obtained by estimating the radius or even optimizing the choice of the center of the ball (cf. the procedure described by Schölkopf et al. (1995); Burges (1998)). However in order to get a theorem of the above form in that case, a more complex argument is necessary: see (Shawe-Taylor et al., 1998, Section VI) for an indication.

We conclude this section by noting that a straightforward extension of the $\nu$-SVC algorithm is to include parametric models $\zeta_k(\mathbf{x})$ for the margin, and thus to use $\sum_k \rho_k \zeta_k(\mathbf{x}_i)$ instead of $\rho$ in the constraint (53) — in complete analogy to the regression case discussed in section 5.

# 8 Experiments

## 8.1 Regression Estimation

In the experiments, we used the optimizer LOQO.[7] This has the serendipitous advantage that the primal variables $b$ and $\varepsilon$ can be recovered as the dual variables of the Wolfe dual (24) (i.e. the double dual variables) fed into the optimizer.

**Toy examples.** The first task was to estimate a noisy sinc function, given $\ell$ examples $(x_i, y_i)$, with $x_i$ drawn uniformly from $[-3, 3]$, and $y_i =$

[7] http://www.princeton.edu/~rvdb/

$\sin(\pi x_i)/(\pi x_i) + \upsilon_i$, where the $\upsilon_i$ were drawn from a Gaussian with zero mean and variance $\sigma^2$. We used the RBF kernel $k(x, x') = \exp(-|x-x'|^2)$, and, unless stated otherwise, $\ell = 50, C = 100, \nu = 0.2, \sigma = 0.2$. Whenever standard deviation error bars are given, the results were obtained from 100 trials. Finally, the **risk** (or test error) of a regression estimate $f$ was computed with respect to the sinc function without noise, as $\frac{1}{6}\int_{-3}^{3} |f(x) - \sin(\pi x)/(\pi x)|\, dx$. Results are given in Table 2 and Figures 3 and 4.

**Table 2** The $\varepsilon$ found by $\nu$-SV regression is largely independent of the sample size $\ell$. The fraction of SVs and the fraction of errors approach $\nu = 0.2$ from above and below, respectively, as the number of training examples $\ell$ increases (cf. Proposition 1).

| $\ell$ | 10 | 50 | 100 | 200 | 500 | 1000 | 1500 | 2000 |
|---|---|---|---|---|---|---|---|---|
| $\varepsilon$ | 0.27 | 0.22 | 0.23 | 0.25 | 0.26 | 0.26 | 0.26 | 0.26 |
| fraction of errors | 0.00 | 0.10 | 0.14 | 0.18 | 0.19 | 0.20 | 0.20 | 0.20 |
| fraction of SVs | 0.40 | 0.28 | 0.24 | 0.23 | 0.21 | 0.21 | 0.20 | 0.20 |

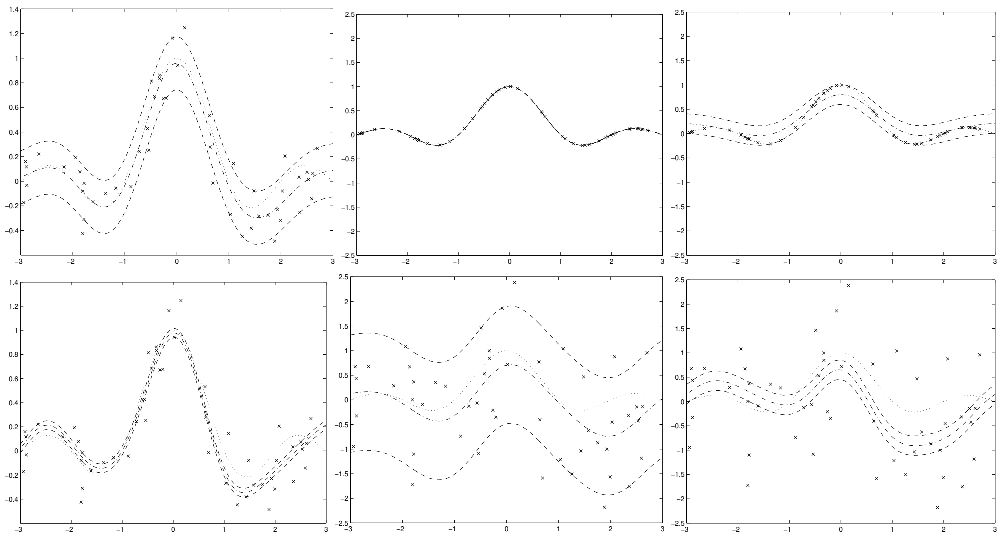

**Figure 3** *Left:* $\nu$-SV regression with $\nu = 0.2$ (top) and $\nu = 0.8$ (bottom). The larger $\nu$ allows more points to lie outside the tube (see section 2). The algorithm automatically adjusts $\varepsilon$ to 0.22 (top) and 0.04 (bottom). Shown are the sinc function (dotted), the regression $f$ and the tube $f \pm \varepsilon$. *Middle:* $\nu$-SV regression on data with noise $\sigma = 0$ (top) and $\sigma = 1$ (bottom). In both cases, $\nu = 0.2$. The tube width automatically adjusts to the noise (top: $\varepsilon = 0$, bottom: $\varepsilon = 1.19$). *Right:* $\varepsilon$-SV regression (Vapnik, 1995) on data with noise $\sigma = 0$ (top) and $\sigma = 1$ (bottom). In both cases, $\varepsilon = 0.2$ — this choice, which has to be specified a priori, is ideal for neither case: in the top figure, the regression estimate is biased; in the bottom figure, $\varepsilon$ does not match the external noise (Smola et al., 1998a).

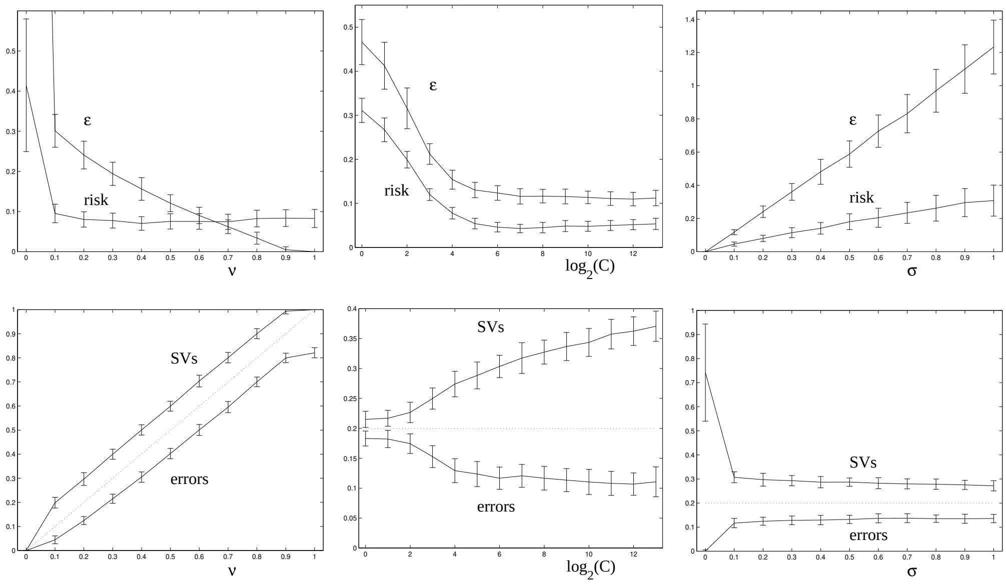


**Figure 4** *Left:* $\nu$-SVR for different values of the error constant $\nu$. Notice how $\varepsilon$ decreases when more errors are allowed (large $\nu$). *Middle:* $\nu$-SVR for different values of the constant $C$. The top figure shows that $\varepsilon$ decreases when the regularization is decreased (large $C$). Only very little, if any, overfitting occurs. In the bottom figure, note that $\nu$ upper bounds the number of errors, and lower bounds the number of SVs (cf. Proposition 1). The bound gets looser as $C$ increases — this corresponds to a smaller number of examples $\ell$ relative to $C$ (cf. Table 2). *Right:* $\nu$-SVR for different values of the noise $\sigma$. The tube radius $\varepsilon$ increases linearly with $\sigma$ (largely due to the fact that both $\varepsilon$ and the $\xi_i^{(*)}$ enter the cost function linearly). Due to the automatic adaptation of $\varepsilon$, the number of SVs and of points outside the tube (errors) is, except for the noise-free case $\sigma = 0$, largely independent of $\sigma$.

Figure 5 gives an illustration of how one can make use of parametric insensitivity models as proposed in section 5. Using the proper model, the estimate gets much better. In the parametric case, we used $\nu = 0.1$ and $\zeta(x) = \sin^2((2\pi/3)x)$, which, due to $\int \zeta(x)\, dP(x) = 1/2$, corresponds to our standard choice $\nu = 0.2$ in $\nu$-SVR (cf. Proposition 5). Although this relies on the assumption that the SVs are uniformly distributed, the experimental findings are consistent with the asymptotics predicted theoretically: for $\ell = 200$, we got 0.24 and 0.19 for the fraction of SVs and errors, respectively.

**Boston housing benchmark.** Empirical studies using $\varepsilon$-SVR have reported excellent performance on the widely used Boston housing regression benchmark set (Stitson et al., 1999). Due to Proposition 2, the only difference between $\nu$-SVR and standard $\varepsilon$-SVR lies in the fact that different parameters, $\varepsilon$ vs. $\nu$, have to be specified a priori. Accordingly, the goal of the following experiment was not to show that $\nu$-SVR is *better* than $\varepsilon$-SVR, but that $\nu$ is a useful parameter to select. Consequently, we are in this experiment only interested in $\nu$ and $\varepsilon$, and hence kept the remaining parameters fixed: we adjusted

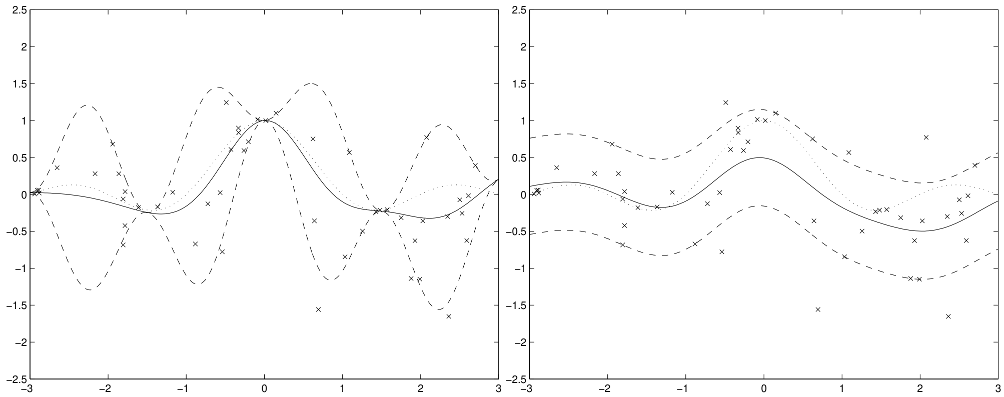


**Figure 5** Toy example, using prior knowledge about an $x$-dependence of the noise. Additive noise ($\sigma = 1$) was multiplied by the function $\sin^2((2\pi/3)x)$. *Left:* the *same* function was used as $\zeta$ as a parametric insensitivity tube (section 5). *Right:* $\nu$-SVR with standard tube.

$C$ and the width $2\sigma^2$ in $k(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2/(2\sigma^2))$ as in (Schölkopf et al., 1997): we used $2\sigma^2 = 0.3 \cdot N$, where $N = 13$ is the input dimensionality, and $C/\ell = 10 \cdot 50$ (i.e. the original value of 10 was corrected since in the present case, the maximal $y$-value is 50 rather than 1). We performed 100 runs, where each time the overall set of 506 examples was randomly split into a training set of $\ell = 481$ examples and a test set of 25 examples (cf. Stitson et al., 1999). Table 3 shows that over a wide range of $\nu$ (note that only $0 \le \nu \le 1$ makes sense), we obtained performances which are close to the *best* performances that can be achieved by selecting $\varepsilon$ a priori by looking at the test set. Finally, note that although we did not use validation techniques to select the optimal values for $C$ and $2\sigma^2$, the performances are state of the art (Stitson et al. (1999) report an MSE of 7.6 for $\varepsilon$-SVR using ANOVA kernels, and 11.7 for Bagging trees). Table 3 moreover shows that also in this real-world application, $\nu$ can be used to control the fraction of SVs/errors.

## 8.2 Classification

As in the regression case, the difference between $C$-SVC and $\nu$-SVC lies in the fact that we have to select a different parameter a priori. If we are able to do this well, we obtain identical performances. In other words: $\nu$-SVC could be utilized to reproduce the excellent results obtained on various data sets using $C$-SVC (for an overview, see Schölkopf et al., 1999). This would certainly be a worthwhile project; however, in the current essay we restrict ourselves to showing some toy examples illustrating the influence of $\nu$ (Fig. 6).

# 9 Discussion

We have presented a new class of SV algorithms, which are parameterized by a quantity $\nu$ which lets one control the number of SVs and errors. We

**Table 3** Results for the Boston housing benchmark; *top:* $\nu$-SVR, *bottom:* $\varepsilon$-SVR. MSE: Mean squared errors, STD: standard deviations thereof (100 trials), Errors: fraction of training points outside the tube, SVs: fraction of training points which are SVs.

| $\nu$ | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| automatic $\varepsilon$ | 2.6 | 1.7 | 1.2 | 0.8 | 0.6 | 0.3 | 0.0 | 0.0 | 0.0 | 0.0 |
| MSE | 9.4 | 8.7 | 9.3 | 9.5 | 10.0 | 10.6 | 11.3 | 11.3 | 11.3 | 11.3 |
| STD | 6.4 | 6.8 | 7.6 | 7.9 | 8.4 | 9.0 | 9.6 | 9.5 | 9.5 | 9.5 |
| Errors | 0.0 | 0.1 | 0.2 | 0.2 | 0.3 | 0.4 | 0.5 | 0.5 | 0.5 | 0.5 |
| SVs | 0.3 | 0.4 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 | 1.0 | 1.0 | 1.0 |

| $\varepsilon$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MSE | 11.3 | 9.5 | 8.8 | 9.7 | 11.2 | 13.1 | 15.6 | 18.2 | 22.1 | 27.0 | 34.3 |
| STD | 9.5 | 7.7 | 6.8 | 6.2 | 6.3 | 6.0 | 6.1 | 6.2 | 6.6 | 7.3 | 8.4 |
| Errors | 0.5 | 0.2 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| SVs | 1.0 | 0.6 | 0.4 | 0.3 | 0.2 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |

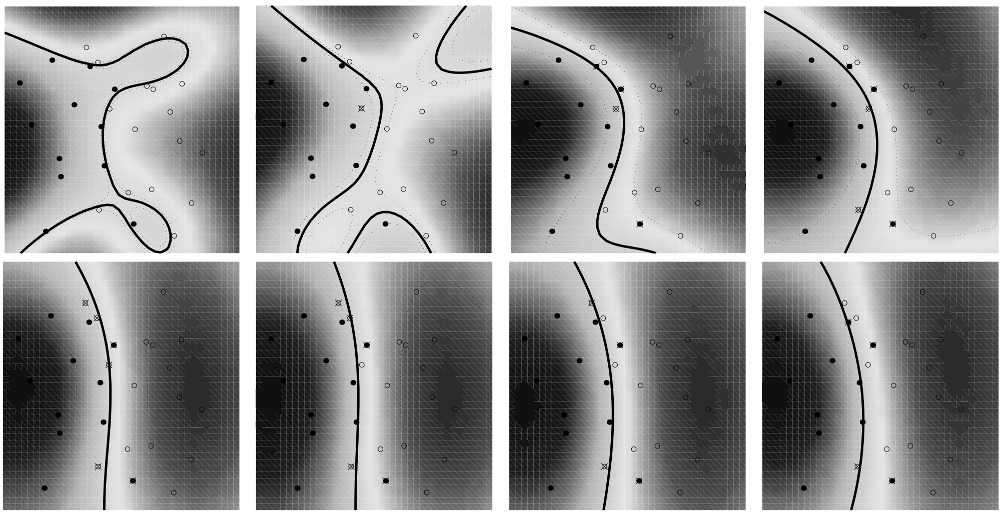

**Figure 6** Toy problem (separate circles from disks) solved using $\nu$-SV classification, using parameter values ranging from $\nu = 0.1$ (top left) to $\nu = 0.8$ (bottom right). The larger we select $\nu$, the more points are allowed to lie inside the margin (depicted by dotted lines). As a kernel, we used the Gaussian $k(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2)$.

described $\nu$-SVR, a new regression algorithm which has been shown to be rather useful in practice. We gave theoretical results concerning the meaning and the choice of the parameter $\nu$. Moreover, we have applied the idea underlying $\nu$-SV regression to develop a $\nu$-SV classification algorithm. Just as its regression counterpart, the algorithm is interesting both from a practical and a theoretical point of view. Controlling the number of SVs has consequences for (1) run-time complexity, e.g., when using a chunking algorithm (Vapnik, 1979) whose complexity increases with the number of SVs; (2) possible data compression applications — $\nu$ characterizes the compression ratio: it suffices to train the algorithm only on the SVs, leading to the same solution (Schölkopf et al., 1995); and (3) generalization error bounds: the algorithm directly optimizes a quantity on which one can give generalization bounds. These, in turn, could be utilized to perform structural risk minimization over $\nu$. Moreover, asymptotically, $\nu$ directly controls the number of Support Vectors, and the latter can be used to give a leave-one-out generalization bound (Vapnik, 1995).

In both the regression and the pattern recognition case, the introduction of $\nu$ has enabled us to dispose of another parameter. In the regression case, this was the accuracy parameter $\varepsilon$, in pattern recognition, it was $C$. Whether we could have as well abolished $C$ in the regression case is an open problem.

Note that the algorithms are not fundamentally different from previous SV algorithms — in fact, we showed that for certain parameter settings, the results coincide. Nevertheless, we believe there are practical applications where it is more *convenient* to specify a fraction of points that is allowed to become errors, rather than quantities which are either hard do adjust a priori (such as the accuracy $\varepsilon$) or which do not have an intuitive interpretation (such as $C$). On the other had, *desirable* properties of previous SV algorithms, including the formulation as a definite quadratic program, and the sparse SV representation of the solution, are retained. We are optimistic that in many applications, the new algorithms will prove to be quite robust. Among these should be the reduced set algorithm of Osuna and Girosi (1998), which approximates the SV pattern recognition decision surface by $\varepsilon$-SVR. Here, $\nu$-SVR should give a direct handle on the desired speed-up.

Future work includes the experimental test of the asymptotic predictions of section 4, and an experimental evaluation of $\nu$-SV classification on real-world problems. Moreover, the formulation of efficient chunking algorithms for the $\nu$-SV case should be studied (cf. Platt, 1999). Finally, the additional freedom to use parametric error models has not been exploited yet. We expect that this new capability of the algorithms could be very useful in situations where the noise is heteroscedastic, such as in many problems of financial data analysis, and general time series analysis applications (Müller et al., 1999; Mattera and Haykin, 1999, e.g.). If a priori knowledge about the noise is available, it can be incorporated into an error model $\zeta$, if not, we can try to estimate the model directly from the data, e.g. by using a variance estimator (e.g. Seifert et al., 1993).

**Acknowledgements** This work was supported in part by grants of the Australian Research Council and the DFG (Ja 379/71 and Ja 379/51). Thanks to K. Müller, J. Platt, R. von Sachs, V. Vapnik and J. Weston for discussions. Jason Weston has independently performed experiments using the sum inequality constraint on the Lagrange multipliers, but declined an offer of co-authorship.

# References

M. Aizerman, E. Braverman, and L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821 – 837, 1964.

P. L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998.

D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, MA, 1995.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, PA, 1992. ACM Press.

C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):1–47, 1998.

C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20: 273 – 297, 1995.

F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10(6):1455–1480, 1998.

R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, Cambridge, 1985.

P. J. Huber. *Robust Statistics.* John Wiley and Sons, New York, 1981.

D. Mattera and S. Haykin. Support vector machines for dynamic reconstruction of a chaotic system. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 211 – 241. MIT Press, Cambridge, MA, 1999.

K.-R. Müller, A. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik. Predicting time series with support vector machines. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 243 – 253. MIT Press, Cambridge, MA, 1999.

E. Osuna and F. Girosi. Reducing run-time complexity in SVMs. In *Proceedings of the 14th Int'l Conf. on Pattern Recognition, Brisbane, Australia*, 1998.

J. Platt. Fast training of SVMs using sequential minimal optimization. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 185 – 208. MIT Press, Cambridge, MA, 1999.

B. Schölkopf, C. Burges, and V. Vapnik. Extracting support data for a given task. In U. M. Fayyad and R. Uthurusamy, editors, *Proceedings, First International Conference on Knowledge Discovery & Data Mining*. AAAI Press, Menlo Park, CA, 1995.

B. Schölkopf. *Support Vector Learning*. R. Oldenbourg Verlag, Munich, 1997.

B. Schölkopf, P. Bartlett, A. Smola, and R. Williamson. Shrinking the tube: a new support vector regression algorithm. In *Advances in Neural Information Processing Systems 11*, 1998a. To appear.

B. Schölkopf, P. Bartlett, A. Smola, and R. Williamson. Support vector regression with automatic accuracy control. In L. Niklasson, M. Bodén, and T. Ziemke, editors, *Proceedings of the 8th International Conference on Artificial Neural Networks*, Perspectives in Neural Computing, pages 111 – 116, Berlin, 1998b. Springer Verlag.

B. Schölkopf, C. J. C. Burges, and A. J. Smola. *Advances in Kernel Methods — Support Vector Learning*. MIT Press, Cambridge, MA, 1999. To appear in December 1998.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299 – 1319, 1998c.

B. Schölkopf, K. Sung, C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik. Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE Trans. Sign. Processing*, 45:2758 – 2765, 1997.

B. Seifert, T. Gasser, and A. Wolf. Nonparametric estimation of residual variance revisited. *Biometrika*, 80:373 – 383, 1993.

J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44:1926–1940, 1998. Also: NeuroCOLT Technical Report NC-TR-96-053, 1996, ftp://ftp.dcs.rhbnc.ac.uk/pub/neurocolt/tech_reports.

J. Shawe-Taylor and N. Cristianini. Robust bounds on the generalization from the margin distribution. In *4th European Conference on Computational Learning Theory*, 1999. To appear.

A. Smola, N. Murata, B. Schölkopf, and K.-R. Müller. Asymptotically optimal choice of $\varepsilon$-loss for support vector machines. In L. Niklasson, M. Bodén, and T. Ziemke, editors, *Proceedings of the 8th International Conference on Artificial Neural Networks*, Perspectives in Neural Computing, pages 105 – 110, Berlin, 1998a. Springer Verlag.

A. Smola and B. Schölkopf. On a kernel-based method for pattern recognition, regression, approximation and operator inversion. *Algorithmica*, 22:211 – 231, 1998.

A. Smola, B. Schölkopf, and K.-R. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637–649, 1998b.

A. J. Smola, S. Mika, and B. Schölkopf. Quantization functionals and regularized principal manifolds. Technical report, Royal Holloway College, 1998c. NeuroColt2-TR 1998-028.

A.J. Smola, T. Frieß, and B. Schölkopf. Semiparametric support vector and linear programming machines. In *Advances in Neural Information Processing Systems, 11*. MIT Press, 1998d. in press.

M. Stitson, A. Gammerman, V. Vapnik, V. Vovk, C. Watkins, and J. Weston. Support vector regression with ANOVA decomposition kernels. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 285 – 291. MIT Press, Cambridge, MA, 1999. to appear.

V. Vapnik. *Estimation of Dependences Based on Empirical Data [in Russian]*. Nauka, Moscow, 1979. (English translation: Springer Verlag, New York, 1982).

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition [in Russian]*. Nauka, Moscow, 1974. (German Translation: W. Wapnik & A. Tscherwonenkis, *Theorie der Zeichenerkennung*, Akademie-Verlag, Berlin, 1979).

G. Wahba. Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 69 – 88. MIT Press, Cambridge, MA, 1999.

R. C. Williamson, A. J. Smola, and B. Schölkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. Technical Report NC-TR-98-019, Royal Holloway College, University of London, UK, 1998.